Telecom Churn Analysis

Javier Lopez

D208, Linear Regression Modeling

Table of Contents

Part I. Research Question

A1. Research Question

What variables impact the amount of bandwidth a customer uses per year?

A2. Goals

Our goal is to understand the relationship between Bandwidth_GB_Year and the explanatory variables. We seek to understand these variables deeply to create a model that will help us predict, with some measure of confidence, the customer bandwidth usage per year using only a fraction of the explanatory variables available. The predictions will help stakeholders make business decisions regarding infrastructure and expansion.

Part II. Method Justification

B1. Assumptions of a Multiple Linear Regression Model

Multiple linear regression models make several assumptions. The first assumption is linearity, and the model assumes that the relationship between the explanatory and target variables is linear. The model also assumes observation is independent of the others and that the data has a normal distribution and no influential outliers. Furthermore, the multiple linear regression model assumes no high correlation between the explanatory variables, also known as multicollinearity. Lastly, there is the

assumption of homoscedasticity. The assumption is that the variance of the residuals is constant across all levels of the explanatory variables.

B2. Benefits of Using Python

Python is a versatile programming language that offers many benefits through data analytics. Python offers benefits in data preprocessing through multiple libraries, including Pandas, Numpy, Statsmodels, and Sklearn. We were able to make use of Pandas and Numpy for data cleaning. Sklearn allowed us to encode categorical variables, which is essential for preparing data for multiple linear regression. Furthermore, Python plays a critical role in model development. Using the Statsmodels and Slearn libraries, we created stepwise regression models and quantified the models' performance using the root mean squared error and r-squared formulas.

B3. Justification: Multiple Linear Regression

Multiple linear regression is an appropriate technique for answering the research question: "What variables impact the amount of bandwidth a customer uses per year?" because it allows us to examine and test the relationship between a dependent variable, bandwidth usage, and multiple independent variables, factors that may impact bandwidth usage. By using multiple linear regression, we can test the impact of the independent variables on the dependent variable while controlling the effects of other variables. This process helps us identify the most significant factors that affect bandwidth usage and how these factors relate to each other. Additionally, multiple linear

regression aids us in the creation of predictive models that can help us estimate future bandwidth usage for customers based on their characteristics.

Part III. Data Preparation

C1. Data Cleaning Goals and Steps

Our goal for data cleaning was to rename non-descriptive columns, impute missing values, remove duplicate values, and identify outliers. We began by implementing the info function and renaming items one through eight columns. We identified less meaningful columns that would not explain the target variable, Bandwidth_GB_Year. Some removed variables included those that served as customer account descriptors and survey responses. We then created a copy of our data frame to preserve the original data and checked for missing values using Pandas' “isna” function.

We continued to utilize Pandas’ “duplicated” function to check for duplicate values; there were none. Before checking for outliers, we calculated the z-scores for our data and, once identified, removed any observations with absolute values greater than three; those were considered outliers. We lost just 8.13% of our observations, leaving us with 91.87% of the data for training and testing. Lastly, we merged our original data frame, keeping only the observations from the index values in the z-scores data frame.

```
## View data types
df.info()
## Rename survey columns
```

```
df.rename({
        'Item1':'TimelyResponse',  'Item2':'TimelyFixes',  'Item3':'TimelyReplacements',
        'Item4':'Reliability', 'Item5':'Options', 'Item6':'RespectfulResponse', 'Item7':'CourteousExchange',
        'Item8':'ActiveListening'
        }, axis=1, inplace=True)
## View summary statistics
df.describe()
## Drop less meaningful columns
df = df.drop([
        'CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State', 'County', 'Zip', 'Lat', 'Lng',
         'Area', 'TimeZone', 'Job', 'Marital', 'Gender', 'Churn', 'Email', 'Multiple', 'OnlineSecurity',
         'OnlineBackup', 'DeviceProtection', 'TechSupport', 'PaperlessBilling', 'PaymentMethod',
         'TimelyResponse', 'TimelyFixes', 'TimelyReplacements', 'Reliability', 'Options',
        'RespectfulResponse', 'CourteousExchange', 'ActiveListening'
        ], axis=1)
## Create copy of dataframe
df1 = df.copy()
## Check for missing values
df1.isna().sum()
## Check for duplicate values
df1.duplicated().value_counts()
## Check for outliers
df1.describe()
## Separate object variables
df2 = pd.DataFrame([df1[col] for col in df1.columns if df1[col].dtype != 'object']).transpose()
df2.info()
## Normalize data and exclude outliers
df2 = df2[zscore(df2).abs() < 3]
```

```
## Count outliers

df2.isna().sum()

## Drop outlier values

df2.dropna(inplace=True)

df2.info()

## Measure data loss

lost = ((len(df1) - len(df2))/len(df1))*100

remaining = 100 - lost

print('{}% of data lost\n{}% of data remains'.format(round(lost, 2), remaining))

        8.13% of data lost

        91.87% of data remains

## Combine dataframes

df = df.loc[df2.index]

df1 = df1.loc[df2.index]

## Reset index values

df = df.reset_index(drop=True)

df1 = df1.reset_index(drop=True)

df1.info()
```

C2. Summary Statistics

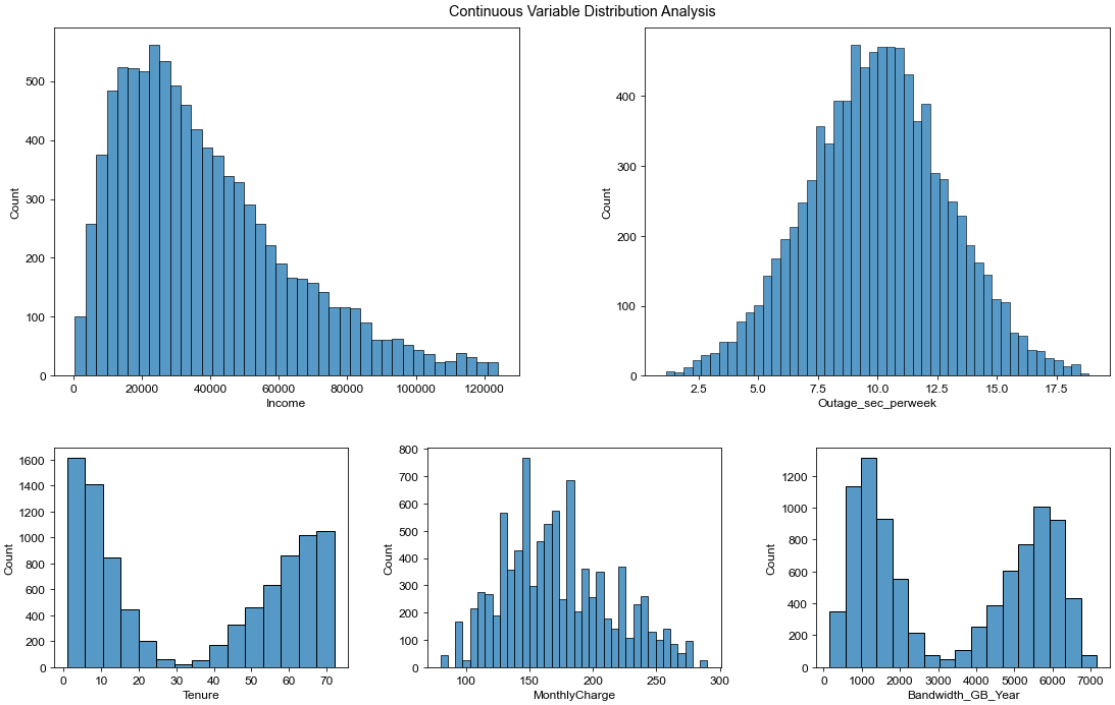
Our dependent variable, Bandwidth_GB_Year, ranges from 155.50 GB (gigabytes) to 7,158.98 GB per year, with an average usage of 3,380.82 GB per year. Most customers, however, are estimated to use just 179.94 GB per year. Our independent variables include a range of data points, including Population, Children,

Age, Tenure, and many others. The table below lists each independent variable's range, standard deviations, mean, median, and mode.

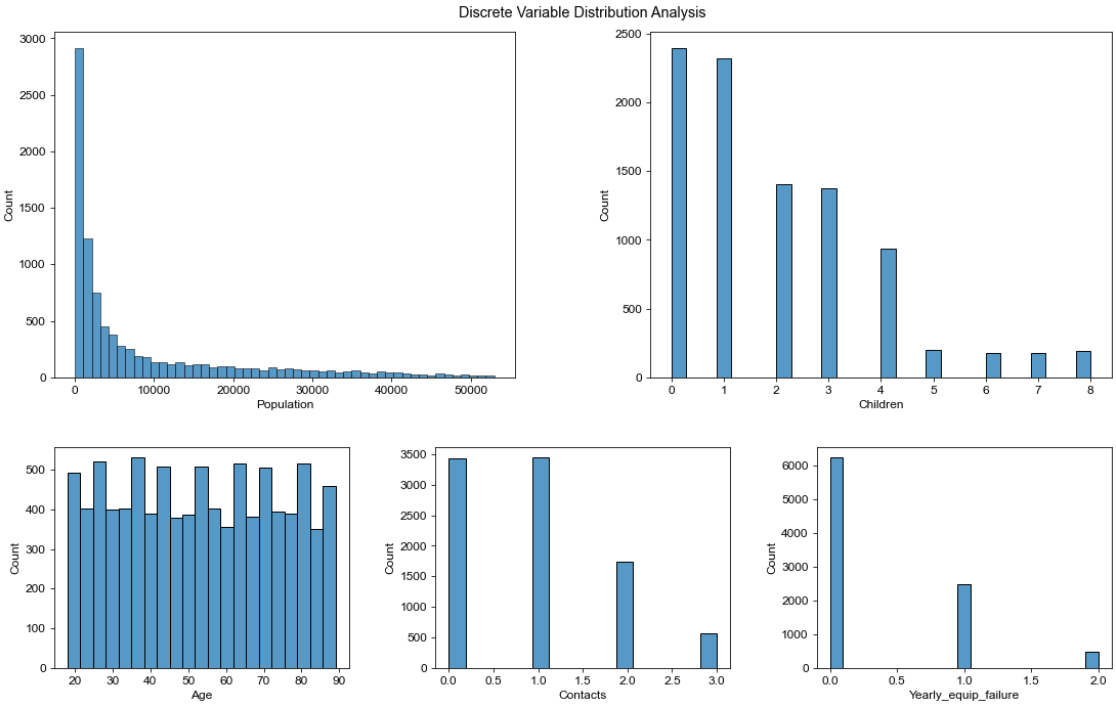
	min	max	std	mean	median	mode
Population	0.000000	52967.000000	11800.643811	8527.956351	2723.000000	0.000000
Children	0.000000	8.000000	1.896481	1.945140	1.000000	0.000000
Age	18.000000	89.000000	20.652026	53.093719	53.000000	63.000000
Income	348.670000	124025.100000	25047.808066	38292.814559	32773.010000	10530.090000
Outage_sec_perweek	1.144796	18.851730	2.927587	10.006093	10.020680	10.488750
Contacts	0.000000	3.000000	0.898659	0.940350	1.000000	1.000000
Yearly_equip_failure	0.000000	2.000000	0.582403	0.374551	0.000000	0.000000
Techie	0.000000	1.000000	0.373826	0.167955	0.000000	0.000000
Contract	0.000000	2.000000	0.835953	0.699793	0.000000	0.000000
Port_modem	0.000000	1.000000	0.499742	0.483945	0.000000	0.000000
Tablet	0.000000	1.000000	0.457441	0.298139	0.000000	0.000000
InternetService	0.000000	1.000000	0.410731	0.785131	1.000000	1.000000
Phone	0.000000	1.000000	0.291893	0.905954	1.000000	1.000000
StreamingTV	0.000000	1.000000	0.499946	0.492653	0.000000	0.000000
StreamingMovies	0.000000	1.000000	0.499901	0.490040	0.000000	0.000000
Tenure	1.005104	71.999280	26.448028	34.443530	30.769800	55.449910
MonthlyCharge	79.978860	290.160419	42.982861	172.733169	169.937800	179.947600
Bandwidth_GB_Year	155.506715	7158.981530	2186.084575	3380.828444	3170.023123	155.506715
InternetDSL	0.000000	1.000000	0.475314	0.344835	0.000000	0.000000
InternetFiberOptic	0.000000	1.000000	0.496423	0.440296	0.000000	0.000000

C3. Distribution Visualizations

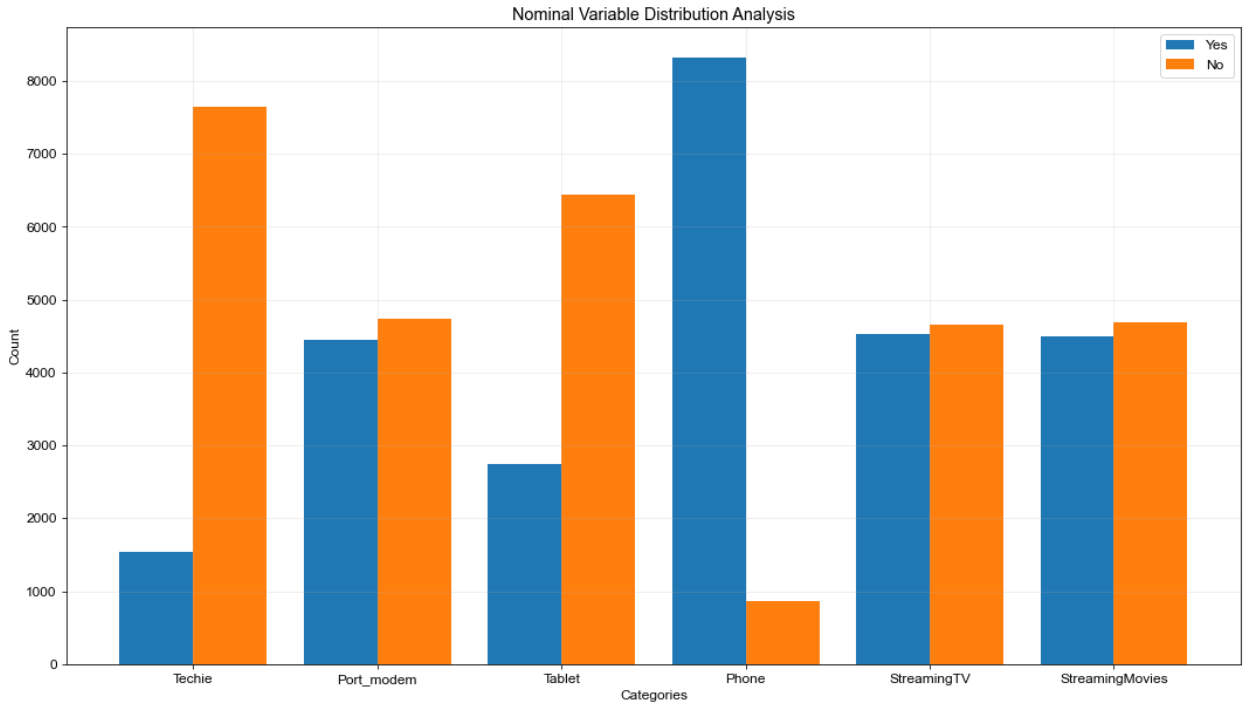
Univariate Visualizations of Continuous Variables

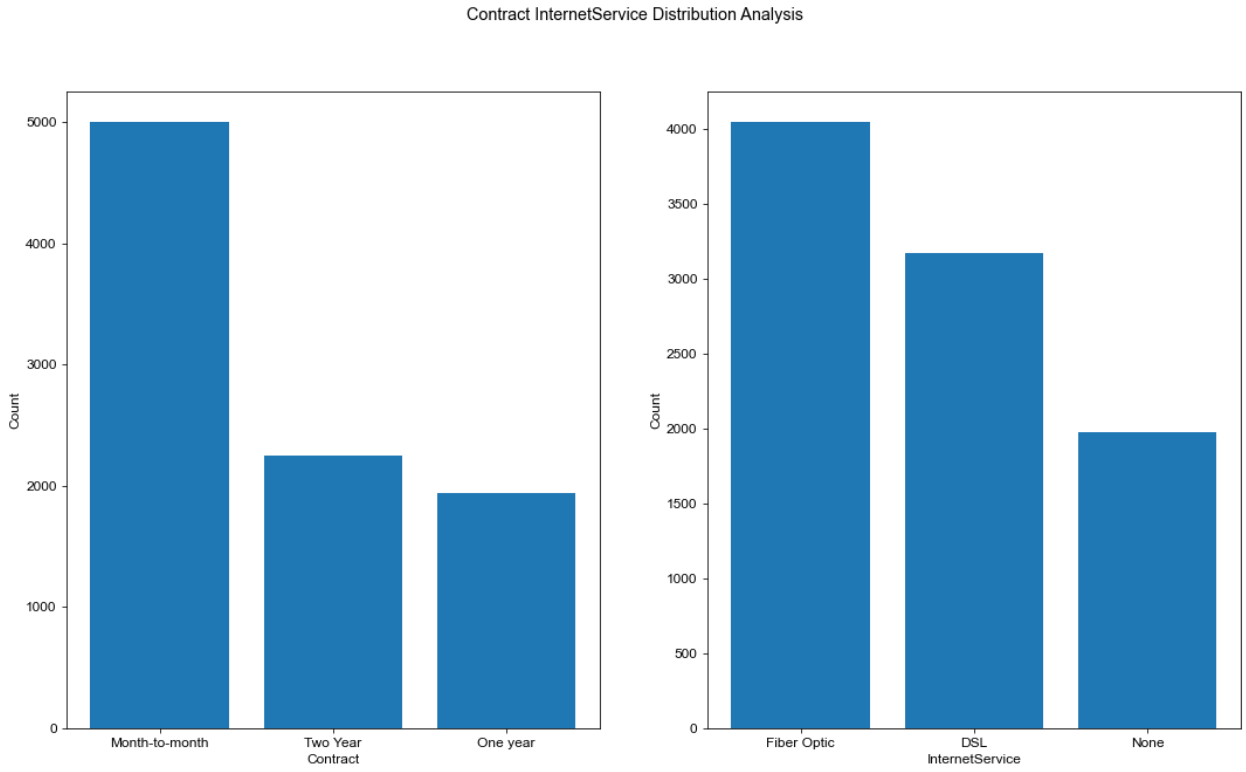


Univariate Visualizations of Discrete Variables

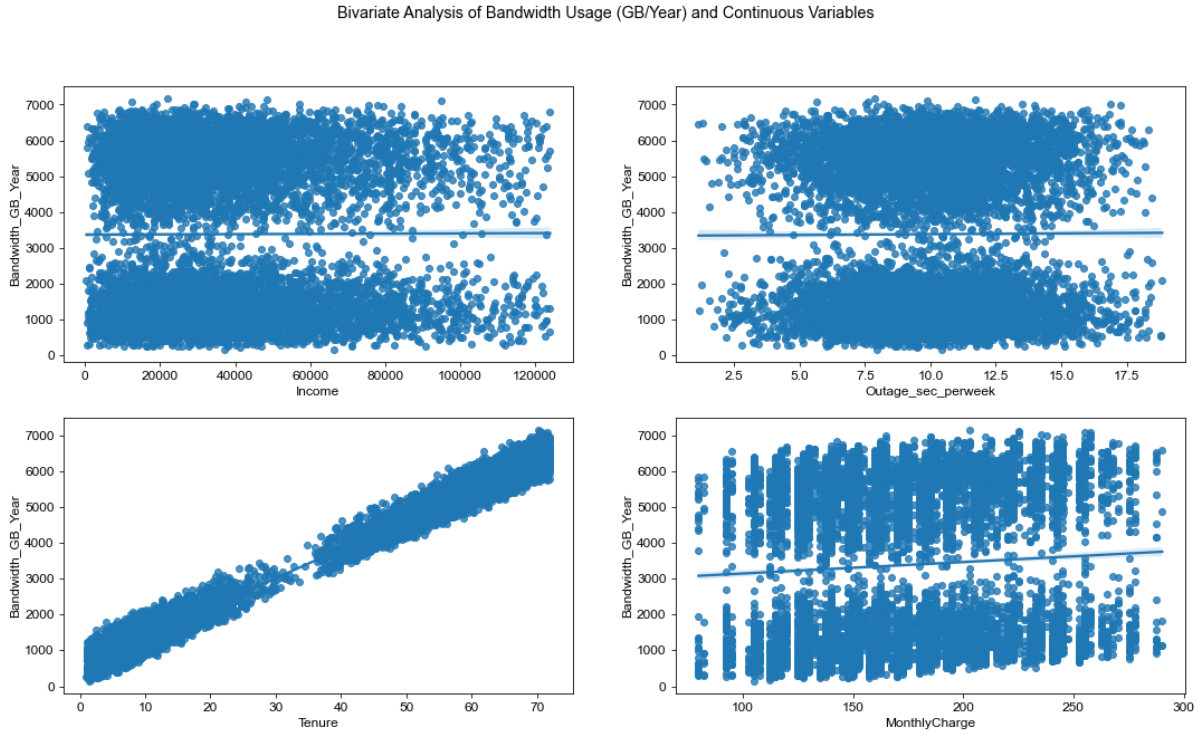


Univariate Visualizations of Nominal Variables

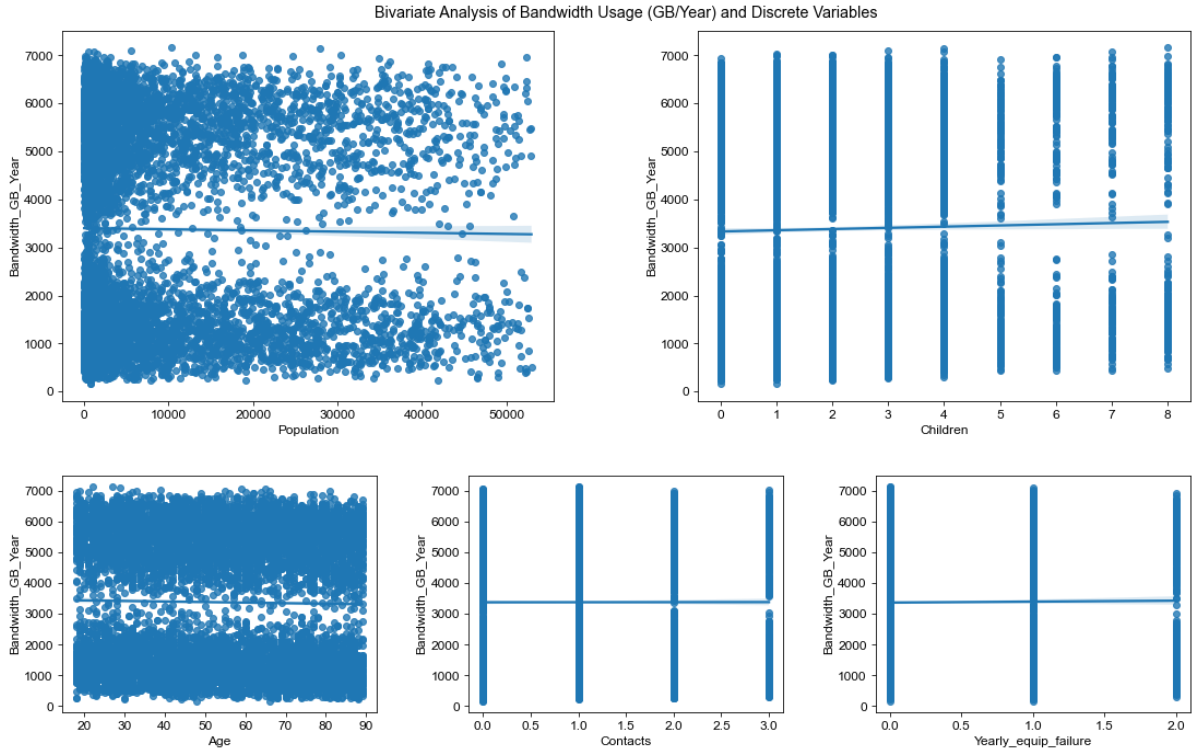




Bivariate Analysis of Continuous Variables



Bivariate Analysis of Discrete Variables



C4. Data Transformation Goals and Steps

Our goals for the data transformation process were to encode all binary nominal variables with one and zero using the label encoder from Sklearn, encode ordinal variables with their respective rank, and create the necessary columns to explain the remainder of our data. We made two columns, DSL and FiberOptic, to describe the type of internet service customers were subscribed to. We used binary to encode the two columns and then transformed the InternetService columns to represent whether a customer subscribed to an Internet service. InternetService was encoded with one if the customer was signed up for DSL or fiber optic internet and zero if the customer was subscribed to neither internet service. We then encoded the contract variable with the respective years the customer signed.

```
## Create InternetDSL and InternetFiberOptic columns
dsl = []
fiber = []
for i in df1.InternetService:
    if i == 'DSL':
        dsl.append(1)
        fiber.append(0)
    elif i == 'Fiber Optic':
        dsl.append(0)
        fiber.append(1)
    else:
        dsl.append(0)
        fiber.append(0)
```

```
df1['InternetDSL'] = dsl

df1['InternetFiberOptic'] = fiber

df1.head()

## Encode InternetService column

internet_service = {'DSL':'Yes', 'Fiber Optic':'Yes', 'None':'No'}

df1.InternetService.replace(internet_service, inplace=True)

## Initiate label encoder

le = LabelEncoder()

## Encode variables

for col in df1.columns:

    if 'Yes' in df1[col].values:

        df1[col] = le.fit_transform(df1[col])

df1.info()

contract = {'Month-to-month':0, 'One year':1, 'Two Year':2}

df1.Contract.replace(contract, inplace=True)

df1.info()
```

C5. Prepared Data CSV

```
## Store clean data as CSV

df1.to_csv('churn_linear_regression.csv')
```

Part IV. Model Comparison and Analysis

D1. Initial Multiple Linear Regression Model

OLS Regression Results

Dep. Variable:	Bandwidth_GB_Year	**R-squared:**	0.999
Model:	OLS	**Adj. R-squared:**	0.999
Method:	Least Squares	**F-statistic:**	5.188e+05
Date:	Mon, 03 Apr 2023	**Prob (F-statistic):**	0.00
Time:	15:32:17	**Log-Likelihood:**	-37887.
No. Observations:	6890	**AIC:**	7.581e+04
Df Residuals:	6871	**BIC:**	7.594e+04
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-16.1294	5.772	-2.795	0.005	-27.444	-4.815
Population	8.419e-06	6.16e-05	0.137	0.891	-0.000	0.000
Children	30.9138	0.374	82.605	0.000	30.180	31.647
Age	-3.2717	0.035	-94.336	0.000	-3.340	-3.204
Income	-3.62e-05	2.83e-05	-1.277	0.201	-9.17e-05	1.94e-05
Outage_sec_perweek	0.1347	0.246	0.547	0.584	-0.348	0.617
Contacts	0.9319	0.795	1.172	0.241	-0.626	2.490
Yearly_equip_failure	-1.5705	1.232	-1.274	0.203	-3.986	0.845
Techie	0.3776	1.915	0.197	0.844	-3.377	4.133
Contract	0.1827	0.856	0.213	0.831	-1.495	1.861
Port_modem	-0.3148	1.429	-0.220	0.826	-3.116	2.486
Tablet	1.2347	1.560	0.792	0.429	-1.823	4.292
InternetService	91.2215	1.254	72.754	0.000	88.764	93.679
Phone	-2.4901	2.465	-1.010	0.312	-7.322	2.341
StreamingTV	98.7062	1.921	51.394	0.000	94.941	102.471
StreamingMovies	52.5620	2.156	24.374	0.000	48.335	56.789
Tenure	81.9097	0.027	3029.588	0.000	81.857	81.963
MonthlyCharge	3.0784	0.031	100.241	0.000	3.018	3.139
InternetDSL	282.5845	1.026	275.313	0.000	280.572	284.597
InternetFiberOptic	-191.3630	1.113	-171.969	0.000	-193.544	-189.182

Omnibus:	129.561	**Durbin-Watson:**	2.012
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	87.514
Skew:	0.156	**Prob(JB):**	9.92e-20
Kurtosis:	2.544	**Cond. No.**	1.08e+19

D2. Justification: Model Reduction Method

Our initial model performed well and explained 99.9% of the variance in our data, but it contained an excessive amount of variables, nineteen variables and a constant, to be exact. To reduce our model, we implemented the use of step-forward feature selection. We began by individually testing each feature in an ordinary least squares model and recording the r-squared value of each model. R-squared measures how much variance in the dependent variable, bandwidth usage, can be explained by the independent variable.

We selected the independent variable with the highest r-squared value and stored the r-squared value and the variable in a dictionary. We then repeated the feature selection process but removed the previously selected variables from the data set. We kept the variables used in each model with their corresponding r-squared value. Once we iterated through all possible independent variables, we sorted the results data frame and selected the model with the lowest root mean squared error. Next, we tested the features chosen for multicollinearity using the variance inflation factor formula from Sklearn and removed any variables with a coefficient above five.

We created a new model using the selected features with low multicollinearity and reviewed the p-values for significance. We removed the variables with a p-value above 0.05 and made our final model using the reduced features. The step-forward feature selection method combined with checking for multicollinearity and significance testing would yield the best model possible using the least possible explanatory variables. Ultimately, our process proved successful, producing a model with a lower RMSE score than the initial model and independent variables with little correlation.

D3. Reduced Multiple Linear Regression Model

Step-Forward Feature Selection Scores and Variance Inflation Factors

Step 1

	RMSE	Features
16	58.453918	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
10	58.699408	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
7	58.95697	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
8	59.052789	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
14	59.155799	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
13	59.241096	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
17	59.520671	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
15	59.73392	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
12	59.799674	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
9	59.978143	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
11	60.13488	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
6	61.751598	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
5	72.045057	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
4	89.770372	[Tenure, InternetDSL, MonthlyCharge, Age, Chil...
3	103.965476	[Tenure, InternetDSL, MonthlyCharge, Age]
2	134.22297	[Tenure, InternetDSL, MonthlyCharge]
1	309.583507	[Tenure, InternetDSL]
0	447.08289	[Tenure]

Step 2

	Variable	VIF
0	Tenure	1.001387
1	InternetDSL	1.327397
2	MonthlyCharge	3.398965
3	Age	1.001600
4	Children	1.002717
5	InternetService	1.523235
6	StreamingTV	1.819095
7	StreamingMovies	2.280884
8	Population	1.001378
9	Income	1.001038
10	Outage_sec_perweek	1.001669
11	Phone	1.001137
12	Contacts	1.001890
13	Yearly_equip_failure	1.000989
14	Port_modem	1.000741
15	Techie	1.000720
16	Contract	1.002108

P-Values for Independent Variables

Step 3

OLS Regression Results

Dep. Variable:	Bandwidth_GB_Year	**R-squared:**	0.999
Model:	OLS	**Adj. R-squared:**	0.999
Method:	Least Squares	**F-statistic:**	5.638e+05
Date:	Mon, 03 Apr 2023	**Prob (F-statistic):**	0.00
Time:	17:21:55	**Log-Likelihood:**	-37822.
No. Observations:	6890	**AIC:**	7.568e+04
Df Residuals:	6872	**BIC:**	7.580e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-13.7374	5.653	-2.430	0.015	-24.820	-2.655
Tenure	81.9305	0.027	3061.435	0.000	81.878	81.983
InternetDSL	475.0420	1.711	277.700	0.000	471.689	478.395
MonthlyCharge	3.0464	0.030	100.277	0.000	2.987	3.106
Age	-3.2795	0.034	-96.005	0.000	-3.346	-3.213
Children	31.3846	0.377	83.260	0.000	30.646	32.124
InternetService	-99.8712	2.122	-47.062	0.000	-104.031	-95.711
StreamingTV	100.9056	1.912	52.775	0.000	97.157	104.654
StreamingMovies	53.7947	2.136	25.180	0.000	49.607	57.983
Population	7.427e-06	5.99e-05	0.124	0.901	-0.000	0.000
Income	-4.197e-05	2.82e-05	-1.488	0.137	-9.73e-05	1.33e-05
Outage_sec_perweek	0.1355	0.242	0.560	0.576	-0.339	0.610
Phone	-3.8351	2.430	-1.578	0.115	-8.599	0.929
Contacts	2.0929	0.792	2.642	0.008	0.540	3.646
Yearly_equip_failure	-2.3894	1.223	-1.954	0.051	-4.786	0.008
Port_modem	1.6147	1.415	1.141	0.254	-1.159	4.388
Techie	1.0671	1.886	0.566	0.572	-2.630	4.764
Contract	-0.3842	0.848	-0.453	0.651	-2.046	1.278

Omnibus:	111.363	**Durbin-Watson:**	1.997
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	77.119
Skew:	0.145	**Prob(JB):**	1.79e-17
Kurtosis:	2.570	**Cond. No.**	3.84e+05

Reduced Model Summary

Step 4

OLS Regression Results

Dep. Variable:	Bandwidth_GB_Year	**R-squared:**	0.999
Model:	OLS	**Adj. R-squared:**	0.999
Method:	Least Squares	**F-statistic:**	1.065e+06
Date:	Mon, 03 Apr 2023	**Prob (F-statistic):**	0.00
Time:	17:21:55	**Log-Likelihood:**	-37827.
No. Observations:	6890	**AIC:**	7.567e+04
Df Residuals:	6880	**BIC:**	7.574e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-17.5951	4.363	-4.033	0.000	-26.148	-9.042
Tenure	81.9291	0.027	3064.786	0.000	81.877	81.982
InternetDSL	474.9970	1.710	277.803	0.000	471.645	478.349
MonthlyCharge	3.0469	0.030	100.300	0.000	2.987	3.106
Age	-3.2795	0.034	-96.012	0.000	-3.346	-3.213
Children	31.3746	0.377	83.300	0.000	30.636	32.113
InternetService	-99.8293	2.122	-47.048	0.000	-103.989	-95.670
StreamingTV	100.9108	1.912	52.774	0.000	97.162	104.659
StreamingMovies	53.7440	2.136	25.159	0.000	49.556	57.931
Contacts	2.0823	0.792	2.629	0.009	0.530	3.635

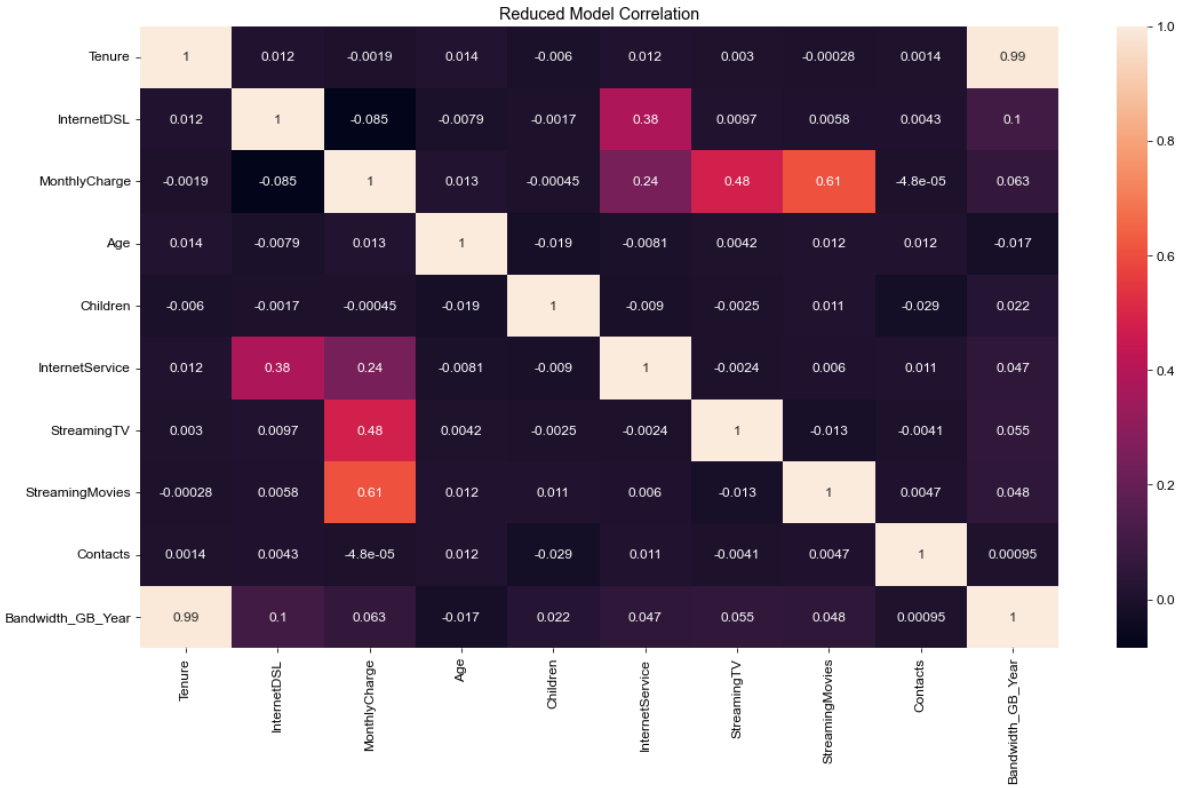
Omnibus:	113.173	**Durbin-Watson:**	1.999
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	77.581
Skew:	0.142	**Prob(JB):**	1.42e-17
Kurtosis:	2.565	**Cond. No.**	1.24e+03

E1. Data Analysis Process

Our data analysis process consisted of cleaning and preparing our data, creating an initial model, step-forward feature selection, feature variance reduction, and testing for feature significance. We began with an initial model that took in data for twenty independent variables and rendered an RMSE (root mean squared error) of 60.01 GB and explained 99.9% of the variance in our data. After following the steps listed above, we produced a reduced model that took in data for just nine independent variables and rendered an RMSE of 59.95 GB while still explaining 99.9% of the variance in our data. Below is a detailed description of the RMSE scores and features used in our reduced model and a heatmap representing the correlation between the independent variable.

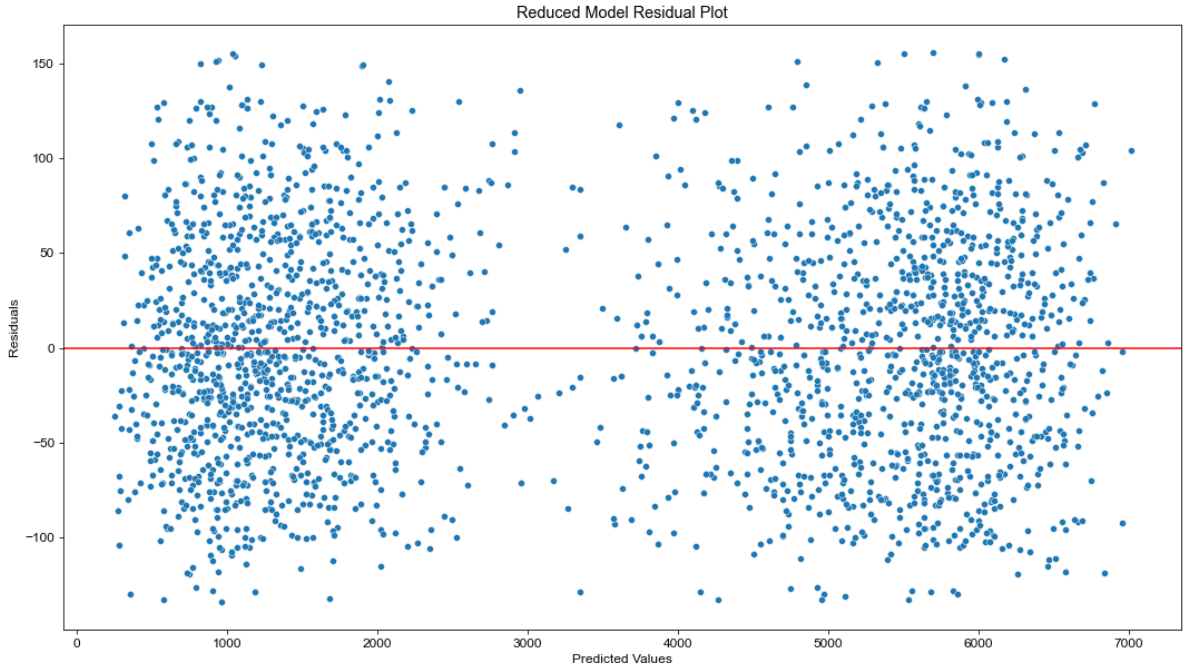
```
Initial RMSE: 60.0124660936566
Initial Features: 20
Final RMSE: 59.954610485256914
Final Features: 9

Final features used:
Tenure
InternetDSL
MonthlyCharge
Age
Children
InternetService
StreamingTV
StreamingMovies
Contacts
```



E2. Results Analysis

Residual Plot



Residual Standard Error

To get the residual standard error, we performed the following calculations:

$$\sqrt{((sum(residuals^2)/ (length(dependent\ variable) - (count\ of\ independent\ variables\ -1)))}$$

After performing the calculation above, we got a total residual standard error of 29.99 GB.

E3. Executable Python File

Executable Python file is saved as: *linear_regression_models.py*.

Part V. Data Summary and Implications

F1. Data Analysis Results

Reduced Model's Regression Equation

The equation for the reduced regression model is:

$$y = intercept + b0 + b1x1 + \ldots + bnxn$$

Where is the dependent variable, b0 is the coefficient of the constant term, b1 and bn are the coefficients for the independent variables, and x1 and xn are the values for the respective independent variable. The equation for our reduced regression would read as follows:

$$Bandwidth\_GB\_Year = - 20.3442 + 81.9112(Tenure) + 473.9745(InternetDSL) + 3.0738(MonthlyCharge) - 3.2426(Age) + 31.1069(Children) - 98.7181(InternetService) + 98.4624(StreamingTV) + 52.3724(StreamingMovies)$$

Coefficient Interpretation

The coefficients of the reduced model tells us how each independent variable affects the dependent variable. For our reduced model, the constant term removes about 20.34 GB to account for any bias in our model. As tenure increases, so does the bandwidth usage of each customer by about 81.91 GB. We can also see that customers subscribed to DSL internet will use about 473.97 GB more than the average customer.

As bandwidth usage increases by about 3.07 GB, a customer's monthly charges are estimated to increase by a dollar. We can also point out that a customer's bandwidth usage will decrease by about 3.24 GB as age increases. As household size increases, so will the customer's bandwidth usage. It is true for our model, as each child is estimated to increase bandwidth usage by about 31.11 GB.

Interestingly, subscribing to internet service is estimated to decrease bandwidth usage by 98.72 GB. However, subscribing to TV streaming will increase usage by 98.46 GB, and movie streaming will increase usage by 52.37 GB. Ultimately, our reduced model takes in a combination of continuous, discrete, and nominal variables. We can use these variables from the regression equation to estimate how much bandwidth a customer consumes. Alternatively, we could predict future usage or run specific scenarios if any independent variables were to change.

Statistical and Practical Significance

Regarding statistical significance, we can use the p-values of each independent variable in our reduced multiple regression model to measure whether the relationship between the independent variable and the dependent is due to chance. Our model summary shows the p-value for all independent variables in our reduced model is

around 0.00. Such a small p-value shows that the relationship is statistically significant, and the probability of observing the relationship by chance is low. We also used the model's R-squared value to determine the statistical significance of the model overall. With an R-squared value of 0.999, we can feel confident that our model explains 99.9% of the variance in our data.

The practical significance of our model is that a telecom company will be able to use the model for multiple prediction purposes. For example, a customer subscribing to streaming services significantly increased bandwidth usage. Those that use more bandwidth have an average increase of one dollar per month for every 3.07 GB of bandwidth they use yearly. A telecom company can use insights like those mentioned above to decide on future marketing strategies and customer incentives.

Limitations

Despite having a statistically significant prediction power and using few data points to make those predictions, there are some limitations to a multiple regression model. One limitation is the multiple regression model's sensitivity to outliers. All outliers must be addressed in the data preparation process, as they can significantly skew the results. Suppose the data has many outliers, and the analyst drops them altogether. The decision to drop too much of the data can impact the accuracy of the linear regression model, as it needs a sufficient sample size to estimate coefficients and make assumptions about the data distribution. It is also important to note that while a multiple linear regression model can identify associations between variables, it cannot prove causation.

F2. Recommendations

When interpreting the coefficients for our model, we saw that the coefficient for Tenure is positive, indicating that as the length of time a customer is with the company increases, so does their bandwidth usage. We can also note that customers who subscribe to DSL, have more children, and are subscribed to streaming services use more bandwidth. However, as customers get older, their bandwidth usage per year decreases. Based on the data, I recommend that the telecom company focus on customer retention, especially customers with children, higher monthly charges, and subscribing to DSL and streaming services. On the other hand, the company can consider ways to provide better service to older customers that use less bandwidth per year. The telecom company can also use the model to predict bandwidth usage for other customers with similar characteristics, which can inform future marketing and pricing strategies.

Part VI. Demonstration

G. Panopto Video

The Panopto video is linked in the project submission.

H. Code Sources

We did not use third-party code to create this project.

I. Citation Sources

We did not use in-text citations in this project.